A Study on Morphological Analyser for Indian Languages: A Literature Perspective

Jayashree Nair, L. S. Aiswarya(✉), and P. R. Sruthy

Department of Computer Science and Applications, Amrita Vishwa Vidyapeetham, Amritapuri, India
jayashree@am.amrita.edu

Abstract. India is the home to a very large number of languages. The Indian languages are rich in literature and has been studied by native and foreign Linguists. Unlike English, Indian languages are Morphologically rich and follows free word-order. Even though there have been efforts towards building morphological analyser for Malayalam and Sanskrit, until now an efficient one is not available. In order to solve this problem we come up with the study on morphological analyzer in Indian languages. Morphological analyser is a linguistic tool that would generate the morphemes of a given word. These rules are based on Indian language linguistics. This paper gives a brief description of the approach used for morphological analyser. With the development of a Python Package that make use of Rule Based Approach for developing Morphological Analyzer. It mainly focusing on noun and this analyzer can be used for Information Retrieval, search engines, Machine Translation, speech recognizer, Text Processing etc.

Keywords: Morphological analyser · Morphemes · Rule based approach · Linguists · Indian languages

1 Introduction

India is a linguistically rich area which had different languages, caste, beliefs, art forms, cultures, religions etc. and it is the home for an uncountable number of different type of lingual families. There are 18 constitutional languages are there in this country, which are written in 10 different scripts. Different states of the country usually speak a different languages in India. English is known a universal language because it is widely spoken language in all over the world, because of that there is a large scope for translation between English and the Indian languages. There is a huge amount of different word forms are there in Indian language therefore it is known as Morphologically rich language consider an example i.e., in English the word 'Tree' has only one form that is 'Trees' similarly when we consider same word in Malayalam. In Fig. 1, there are different forms of Malayalam words for the English word 'Tree'.

M. Singh et al. (Eds.): ICACDS 2021, CCIS 1440, pp. 112–123, 2021.
https://doi.org/10.1007/978-3-030-81462-5_11

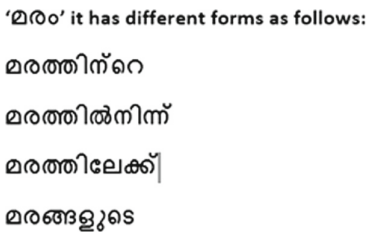

Fig. 1. Different forms of Malayalam word 'maram'

Simple lexical mapping will not help for retrieving and mapping of Morpho-syntactic information from English language. Morphology is a branch of linguistics unit that form a word in Natural Language i.e., it is the field of linguists that are concentrated on the study of formation of words. Every language has a set of words, combining this words along with their grammar with respect to their language to form meaningful sentences. To consider the arrangement of words we need Identification, analysis, and description of structure of a given morphemes of a language likewise to realize how words are worked from more modest part and other linguistic units, for example, root words, suffixes, affixes and so on. A single word in a language is the combination of one or more morphemes and this can be either a root word, suffix or prefix, for example happy, unhappy, happily. Morphological analyser is a linguistic tool that would generate the morphemes of a given word. For morphological analysis most of the NLP system make use of simple linguistic theories. The syntax of Morphological analyzer is:

$$\boldsymbol{Word = stem/root + suffix}$$

The major use of this Morphological Analysers is in search engines, speech recognizer, spell and grammar checker and machine translation [2]. Malayalam is a member of Dravidian Language family because of a highly inflectional and Agglutinative character. This has posed a challenge for all kind of language processing. If we look at the Sanskrit literature, we see that many attempts have been made to render the learning of Sanskrit word formation easier [10]. A dictionary is an arrangement of different words with respect to their meanings, usage, origins, pronunciations etc. For data collection a root dictionary is there which contains root words and its suffix of Sanskrit and Malayalam. Rule based approach is used here, A rule-based approach applies human-made principles to store, sort and control information. In doing as such, it emulates human insight. To work, rule-based system require a bunch of realities or wellspring of information, and a bunch of rules for data manipulation [9]. This paper is the study to build an efficient Morphological Analyzer using rule based approach.

2 Literature Review

This section include different terms in Literature Review such as NLP(Natural language processing), Indian Languages, Malayalam, Sanskrit, Morphological Analyser, Morphology, Root Dictionary.

2.1 Natural Language Processing

Natural language processing (NLP) is both a contemporary computational technology and a way of investigating and evaluating claims about human language itself. Some choose the time period computational linguistics for you to seize this latter characteristic, however NLP is a term that links again into the records of Artificial Intelligence (AI).

Natural Language Processing is an interdisciplinary field where linguistic and computer science merge. To build computational models of natural language for its analysis and generation is the ultimate aim of NLP and is mainly focused on the study of language for communication [7]. It is the process of making the human language more easier and understandable to machines and performing different operations on it to retrieve useful information.

2.2 Indian Languages

India is the place having a large variety of languages, religions, cultures ets. Each languages have different word formation, grammatical features and script. India gain 4 th position in the world for having large number of languages. There are 18 constitutional languages are there in this country, which are written in 10 different scripts. Different states of the country usually speak a different languages in India. Hindi is known as the official language of India.

2.3 Malayalam

Malayalam is one of the 22 scheduled Indian language, over 34 million people spoke this language. It is the official language in Kerala Malayalam language comes under Dravidian family which have characteristic feature of agglutinative language which consisting of 15 vowels known as swarah akshara and 36 consonants which is vyenjana akshara also have symbols like chillu akshara. Malayalam language comprising of free consonant and vowel likewise has its own particular content, a syllabic letters in order, [4].

2.4 Sanskrit

Sanskrit is known as mother of all languages in India because it is the traditional means of communication and is used in ancient poetry, drama, religious and also in philosophical texts. This language is rich inflectional as well as derivational morphology [10]. The writing script used for Sanskrit is Brahmi script and this language is belong to Devanagari language family. There are 46 alphabets are there in this language which contains 16 vowels it is known as swaras.

2.5 Morphology

The study of morphemes is known Morphology, a morpheme means smallest unit of meaning in a language that form a word and we cannot divide further [8]. When we take a meaningful word that must have at least 1 morpheme, and also there are words with many morphemes too. Free and Bound are the two type of morphemes Fig. 2. Free morpheme has a meaning in the language ie independent word whereas bound is dependent word which are meaningless.

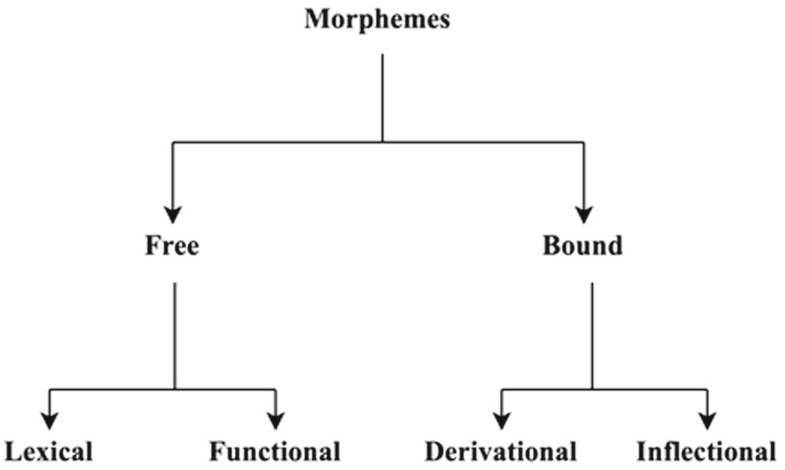

Fig. 2. Types of morpheme

Let us look it with an example Fig. 3, consider the word "Socialist" there are only 2 morphemes Social+ist and this cannot further divide. Here Social is free morpheme and ist is bound morpheme.

2.6 Morphological Analyser

Morphological analyser take a word as input and give an output as grammatical information which include root word and suffix. It is intended to dissect the constituents of the words and it will help for the division of words into stems. The Morphological analyser which return root/stem word alongside its syntactic data relying on its word category [9]. The general format for the morphological analyser is:

$$Word = stem/root + suffix$$

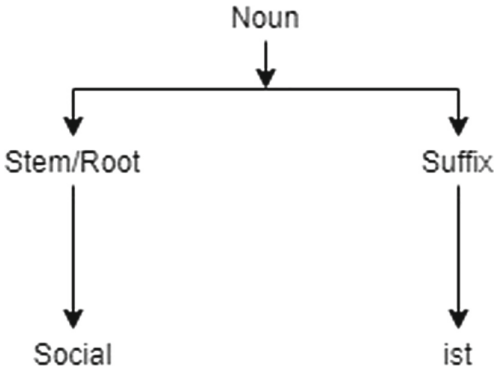

Fig. 3. Example of morpheme

2.7 Root Dictionary

A dictionary is an arrangement of different words with respect to their meanings, usage, origins, pronunciations etc. words are wrapped in a single resource. This root dictionary is the collection of romanized list of root and suffix words.

3 Different Python Packages for Morphological Analysis

Description of all available packages for morphological analyser such as INLTK (Natural Language Toolkit for Indian Languages) and polyglot.

3.1 INLTK (Natural Language Toolkit for Indian Languages)

iNLTK presents assist for numerous NLP applications in indic languages. The languages supported are Hindi denoted as (hi), Punjabi denoted as (pa), Sanskrit denoted as (sa), Gujarati denoted as (gu), Kannada denoted as (kn), Malayalam denoted as (ml), Nepali denoted as (ne), Odia denoted as (or), Marathi denoted as (mr), Bengali denoted as (bn), Tamil denoted as (ta), Urdu denoted as (ur), English denoted as (en). INLTK is similar to the nltk python bundle and that they offers same features for nlp along with tokenisation and vector embedding for enter text with an clean API interface. The Indian languages have some difficulties which come from sharing a lot of similarity in terms of script, phonology, language syntax, etc., and this library provides a general solution. Indic NLP Library provides functionalities like text normalisation, script normalization, tokenization, word segmentation, romanization, indicisation, script conversion, transliteration and translation.

3.2 Polyglot

Polyglot is used to perform different NLP operations and is an open-source python library. Different types of python libraries are available that can help us

in performing NLP assignments. All libraries have certain novel highlights and which make them unique in relation to one another. This polyglot has an enormous variety of dedicated commands which makes it stand apart of the group also computation is depends on NumPy which is the reason it is known as quick and faster. It is similar to spacy and as compare to spacy this polyglot is better i.e., it can be used for languages that are not support by spacy. Polyglot provide different functionalities and can imported as and when required. Polyglot can recognize the language of the content passed to it using language function. In Morphological analysis it characterizes the consistencies behind word arrangement in human language. It can be used to identify the language in a specific text, followed by the tokenization in words and sentences. It is easy to use and can be used for a variety of NLP operations depending on our need.

Polyglot has proved to be a completely beneficial device for experimenting with new language features and for building different language-processing gear. Polyglot isn't only a preprocessor it supports the development of complicated language extensions that add new features to the java language, consisting of to its type device. This library is not a well-known library but it offers a wide range of analysis and splendid language coverage. And this library stands out from the crowd also because it requests the usage of a dedicated command in the command line through the pipeline mechanisms. It also works really fast and also it's very efficient, straightforward, and basically an excellent choice for projects involving a language Spacy doesn't support.

4 Comparative Study of Existing Morphological Analyzer for Indian Languages

A comparative study of different python packages had done which include Spacy, Inltk and Polyglot. In Spacy there is no morphological Analyser for malayalam, and for malayalam they have only limited functionalities like wordnet and lematization. In Inltk there is no morphological analyser instead of that they have stemmer and lemmatization and this stemmer removes only the root word and lemmatization combines words with similar words into one. In polyglot it have morphological analyzer for Malayalam but it doesn't provide an accurate result, for that we had done a test Fig. 4, For the test we had randomly select 100 words with suffix, which contain both nouns and verbs. When we import this file of 100 words in CSV into this polyglot. It will give output by splitting words into its root and suffix and many of them are wrong. we can clearly understand this from this chart. Here 100 words are there among that 37 of them are nouns and in that nouns 21 of them are correct and 16 are wrong, similarly 63 of them are verbs and 42 of them are correct and remaining 21 of them are wrong.In the comparative study we are trying to say that even though we had an analyser it doesn't give an accurate result i.e., for Malayalam until now there is not an efficient morphological analyser. This knowledge will help to develop an efficient morphological analyser for Malayalam.

Input Type	Total count	Correct Result	Error Result
Nouns	37	21	16
Verbs	63	42	21
Result	100	63	37

Fig. 4. comparative study

5 Related Works

Morphological analysis is the division of words into their part morphemes and the task of linguistic data to syntactic classifications. It contain identification of parts of the words, or more technically, constituents of the words and it will return its roots/stem of a word along with its grammatical information depending upon its word category [9]. There are different morphological analyzer developed for various languagaes using different approaches like hybrid approach, Finate state transducer, rule based approach, suffix stripping approach and so on.

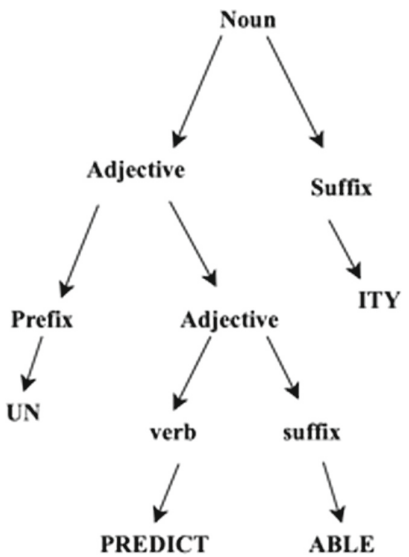

Fig. 5. Morphological structure of a word

Here the word 'Unpredictability' Fig. 5 is formed from an adjective and a suffix ie adjective is 'unpredictable' and suffix is 'ity' then this adjective is formed from a prefix and another adjective ie 'predictable' is the adjective and it is formed from a verb 'predict' and suffix 'able'.

5.1 Hybrid Approach

For Indian languages, hybrid approach which combines the advantages of the rule based approach and Data driven approach. This approach is simple because all the inflections are stored in the database and the inflections can be found directly from the database based on the given word. The morphological inflections of all the words of a Morphologically rich language is highly difficult to pre-record.

5.2 Finite State Transducer

The finite state transducer(FST) is proposed for morphological analyzer for Malayalam language and this FST approach which maps strings from one regular language into strings from another regular language and this process is reversible too [6]. FST is used to represent the lexicon computationally and it can be done by accepting the principle of two level morphology. That is, it which represents a word as the correspondence between lexical and surface level of morphology. An FST is represented as a two tape automation. By combining the lexicon, orthographic rules and spelling variations in the FST we can build a morphological analyzer [11]. The transducers is a kind of translating machine which read from one tape and write into the other. The FST method which acts as a two level morphology and this method is used for both analysis and generation.

5.3 Suffix Stripping

The suffix stripping method is used for the development of the morphological analyzer for Indian languages and it make use of a stem dictionary. This dictionary which contain all possible suffixes that nouns/verbs in the language can have morphotatic rule and morphophonemic rule [4]. This technique which identifies the suffix first and then the stem by way of making use of morphophonemic rules. The suffix stripping approach which is simpler to maintain and the searching process are relatively fast as the search is only done on suffixes. As we know, the Words are formed by adding suffixes to the root words. So this property can be appropriate for suffix stripping method. Once the suffix is identified, the root of the whole word can be acquired by removing the suffix from the root word and applying proper orthographic (sandhi) rules. There are set of dictionaries like root dictionary, suffix dictionary and also using morphotactics and sandhi rules, and they are used in suffix stripping algorithm [12].

5.4 Rule Based Approach

The rule based approach and this approach is generally used for building morphological analyser. This approach is very effective for the morphological analysis of Indian languages and it predicts correct grammatical features. This approach which works based on some set of guidelines and dictionary that includes roots and morphemes [2]. There are set of rules are there in this approach and these rules are directly or indirectly depends on each other.

Here we using rule based approach because this approach is mainly used for building the morphological analyser. In this approach, which include a set rules and dictionary that includes root and morphemes and there rules are both manually created or extracted from a large corpus that based on a few common features. The rules in this approach is both right away or indirectly depends on previous rule i.e., all of the rules which is depends on every different. This is used because, it is straightforward and this make it ideal and less time consuming. For the working of this rule based system requires a set of facts or source of data, and a set of rules for manipulating that data. And it also proves very effective and provides better accuracy than the existing ones.

In Fig. 6 paper [3] 'Morphological Analyzer for Malayalam Using Machine Learning' they have used rule based method for the analyser by using SVM tool and these rules are automatically learned from the data, for learning models and make predictions they used learning and classification algorithms. This gives Efficient output and also it correctly predict the grammatical features of words which are not available in the training set.

'FST Based Morphological Analyser for Hindi Language' is another paper [1] mentioned above, in this paper they utilizes Stuttgart Finite State Transducer (SFST) device for creating FST. A root word dictionary is made here. Rules are then added for creating inflectional and derivational words from these root words. The Morph Analyser created was utilized in a Part Of Speech (POS) Tagger dependent on Stanford POS Tagger. The framework was first prepared utilizing a physically labeled corpus and for labeling input sentences MAXENT (Maximum Entropy) approach of Stanford POS tagger was used.

The paper 'Morphological Analyser for Hindi using Rule Based Approach' [11] which is mentioned above uses the Rule Based approach. It uses lemmatize to extract the root word properly and a corpus which stores the exceptional words which does not match with the rule made. For the development of the corpus, commonly used words are used. Hindi morphological structure which consists of various word classes in which their derivational and inflectional forms are described. The rules are made to comprise almost all the phrase formations available after a deep evaluation and observe of the dictionary and different expertise assets to be had.

The paper 'A Graph Based Semi-Supervised Approach For analysis of derivational Nouns in Sanskrit' uses the Semi-Supervised graph based approach for morpho-syntactic lexicon induction. The Modified Adsorption(MAD) algorithm is used for the task. This approach is used for the analysis of derivational nouns in Sanskrit.

Sl.No	Paper	Method	Language	Contributions	Scope or limitation
1	Morphological Analyzer for Malayalam Using Machine Learning	Rule based method	Malayalam	SVMTool is applied to the Analyzer for predicts right syntactic highlights	Major rules of rule based approach is used because of that if one rule fails it will affect the entire rule that follows.
2	Morphological Analyser and Morphological Generator for Malayalam - Tamil Machine Translation	Suffix Joining and Suffix Stripping method is used.	Malayalam and Tamil	Bilingual word dictionary reference for Malayalam and Tamil comprise of the root/stem of the words with its linguistic class.	Malayalam tends to join two words, this is one of the significant issue.
3	Morphological analyzer and generator for Tulu language: a novel approach	Rule based method	Tulu	In view of the inflections and differences, all conceivable Morphotactic and Sandhi rules were composed.	This framework can be improved by adding more rules, eg: rules for complex morphology, rules for transitive structures.
4	FST Based Morphological Analyzer for Hindi Language	Stuttgart Finite State Transducer approach	Hindi	The Analyzer created was utilized the feature (POS) Tagger dependent on Stanford POS Tagger.	The Morph Analyzer can be enhanced by combining the paradigm approach with the FST approach.
5	A hybrid approach to Tamil morphological generation.	Hybrid approach	Tamil	This Morphological Generation combines the advantages of the rule based approach and Data driven approach.	This algorithm can be used to generate morphological generator for pronouns and adverbs.
6	A Graph Based Semi-Supervised Approach for Analysis of Derivational Nouns in Sanskrit	Graph based Semi-Supervised approach	Sanskrit	MAD(Modified Adsorption) algorithm used for the task.	There exists no analyser for Sanskrit and this prompts issues with tremendous extension getting ready of compositions in Sanskrit.
7	Morphology analysis for Malayalam Language using FST	Finite State Transducer(FST) approach	Malayalam	morphology model is compiled base using HFST toolkit. Ml-morph uses its own POS tagging schema	The collected words are manually assure and clean up because of that this task is tedious, however is incredibly vital to the standard of analyser.
8	Morphological Analyser for Hindi -A Rule Based implementation	Rule Based Approach	Hindi	uses lemmatize to extract the root words properly and a corpus which stores the exceptional words which does not match with the rules made	It will integrate the word sense elucidation with this analyser therefore that the words having multiple senses can be analyzed accurately.

Fig. 6. A consolidated report charting all the important Morphological Analyzer Papers.

6 Research Gaps and Novelty

Morphological analyzer is a linguistic tool that would generate the morphemes of a given word. It is designed to analyse the constituents of the words and it will help for the segmentation of words into stems. The morphological analyzer which go back root/stem phrase in conjunction with its grammatical facts relying upon its phrase class [9]. The general format for the morphological analyzer is:

$$Word = stem/root + suffix$$

Example:of seetha => സീതയുടെ=സീത+ഉടെ

This example give clear view of what this morphological analyzer does in Malayalam language i.e., the Morphological analyzer will split the given input word into its own root and suffix. morphology is the study of word formation, how words are constructed up from smaller portions. i.e., to identification, analysis, and description of the shape of a given language's morphemes and other linguistic gadgets, which include root words, affixes, components of speech or implied context. Analysis of phrase or word structure (morphology) is split into fundamental fields as inflection and derivation. Consequently, the morphological shape of each phrase might also encompass factors such as prefix, suffix, infix, or even a separate root, and those factors can modify the meaning of the simple root or stern of the phrase. If the resultant phrase is only a paradigmatic application of its base shape, this modification of the phrase is known as inflection; however if the ensuing word is a different phrase or a compound, which is formed of two or more roots, it is known as derivation. At the same time as derivation is a word-creating method, inflection constitutes exclusive styles of any phrase or word [8]. For the development of the morphological analysis, rule based approach is mainly used, because it is a system that applies human made rule to store, sort and manipulate data. This approach is based on set of fixed rules and dictionary that contains root and morphemes and the rules which contained in this approach are made to incorporate and other knowledge resources available. The rules are made to comprise almost all the phrase formations available after a deep evaluation and observe of the dictionary and different expertise assets to be had.

6.1 Need of Morphological Analysis

Morphological Analysis is a decent organized technique and approach that assists with finding new connections or arrangements which may be disregarded by other less organized strategies [3]. It is a foundational and centered strategy, which permits coordinating existing data and producing new innovative thoughts for planning new items, advances and services.

7 Conclusion

This paper presents a short study on Morphological Analyzer for Indian Languages which is mainly concentrated to Malayalam and Sanskrit language. These

are morphologically rich as compared with English. Morphological analyzers can be used for Information Retrieval, search engines, Machine Translation, speech recognizer, Text Processing etc. This paper drafts a brief presentation on the existing python packages that include MA. The comparison report on each of them is also depicted.

References

1. Kumar, D., Singh, M., Shukla, S.: FST based morphological analyzer for Hindi language. arXiv preprint arXiv:1207.5409 (2012)
2. Antony, P.J., Raj, H.B., Sahana, B.S., Alvares, D.S., Raj, A.: Morphological analyzer and generator for Tulu language: a novel approach. In: Proceedings of the International Conference on Advances in Computing, Communications and Informatics, pp. 828–834, August 2012
3. Abeera, V.P., et al.: Morphological analyzer for Malayalam using machine learning. In: Kannan, Rajkumar, Andres, Frederic (eds.) ICDEM 2010. LNCS, vol. 6411, pp. 252–254. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-27872-3_38
4. Jayan, J.P., Rajeev, R.R., Rajendran, S.: Morphological analyser and morphological generator for Malayalam-Tamil machine translation. Int. J. Comput. Appl. **13**(8), 0975–8887 (2011)
5. Krishna, A., et al.: A graph based semi-supervised approach for analysis of derivational nouns in Sanskrit. In: Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing, pp. 66–75, August 2017
6. Thottingal, S.: Finite State Transducer based Morphology analysis for Malayalam Language. In: Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages, pp. 1–5 (2019)
7. Chaitanya, V., Sangal, R., Bharati, A.: Natural Language Processing: A Paninian Perspective. Prentice-Hall of India, Delhi (1996)
8. Saranya, S.K.: Morphological analyzer for Malayalam verbs. Unpublished M. Tech Thesis, Amrita School of Engineering, Coimbatore (2008)
9. Jayan, J.P., Rajeev, R.R., Rajendran, S.: Morphological analyser for Malayalam-a comparison of different approaches. IJCSIT **2**(2), 155–160 (2009)
10. Kulkarni, A., Shukl, D.: Sanskrit morphological analyser: some issues. Indian Linguist. **70**(1–4), 169–177 (2009)
11. Agarwal, A., Singh, S.P., Kumar, A., Darbari, H.: Morphological Analyser for Hindi-A rule based implementation. Int. J. Adv. Comput. Res. **4**(1), 19 (2014)
12. Antony, P.J., Soman, K.P.: Computational morphology and natural language parsing for Indian languages: a literature survey. Int. J. Sci. Eng. Res. **3** (2012)